AI: What Is To Be Done?

Stuart Russell
UC Berkeley

What is AI?

AI = making intelligent machines

Standard model: rational behaviour given human-defined objectives

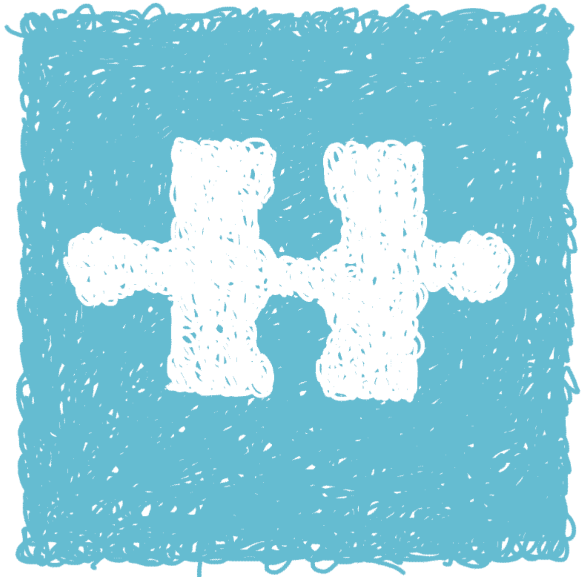
The goal is general-purpose AI: capable of quickly learning high-quality behavior in “any” task environment

Have we succeeded?

No

GPT-like models are probably a piece of the puzzle but as fixed-size feedforward circuits they are fundamentally limited

We don’t know what shape the piece is or where it goes



Doesn’t deep learning solve everything?

Feedforward circuits (including LLMs) are linear-time

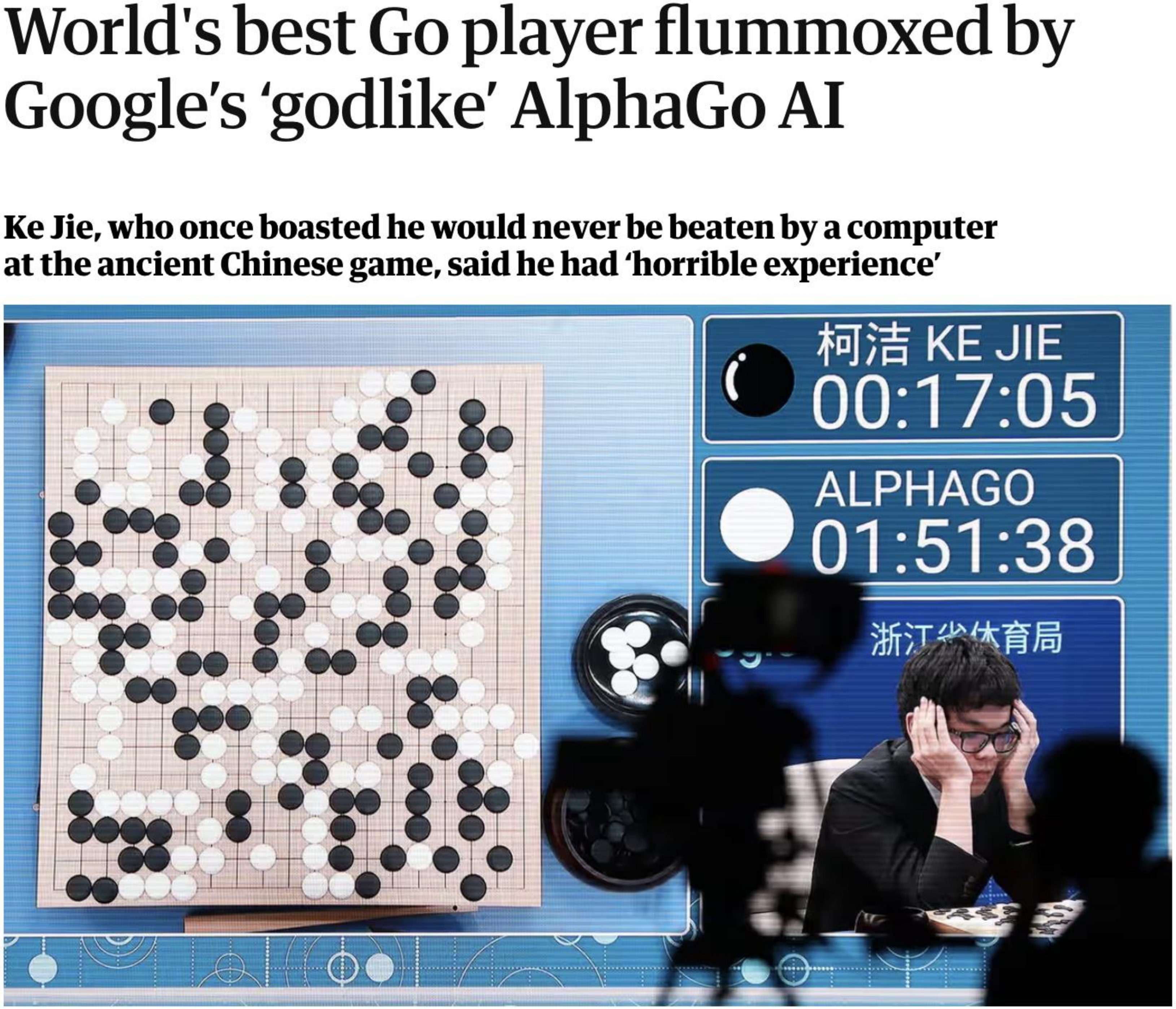
- Amount of computation is exactly proportional to the size of circuit

Many tasks are superlinear (in input size)

- E.g., NP-hard tasks require exponential time

=> Very large circuits

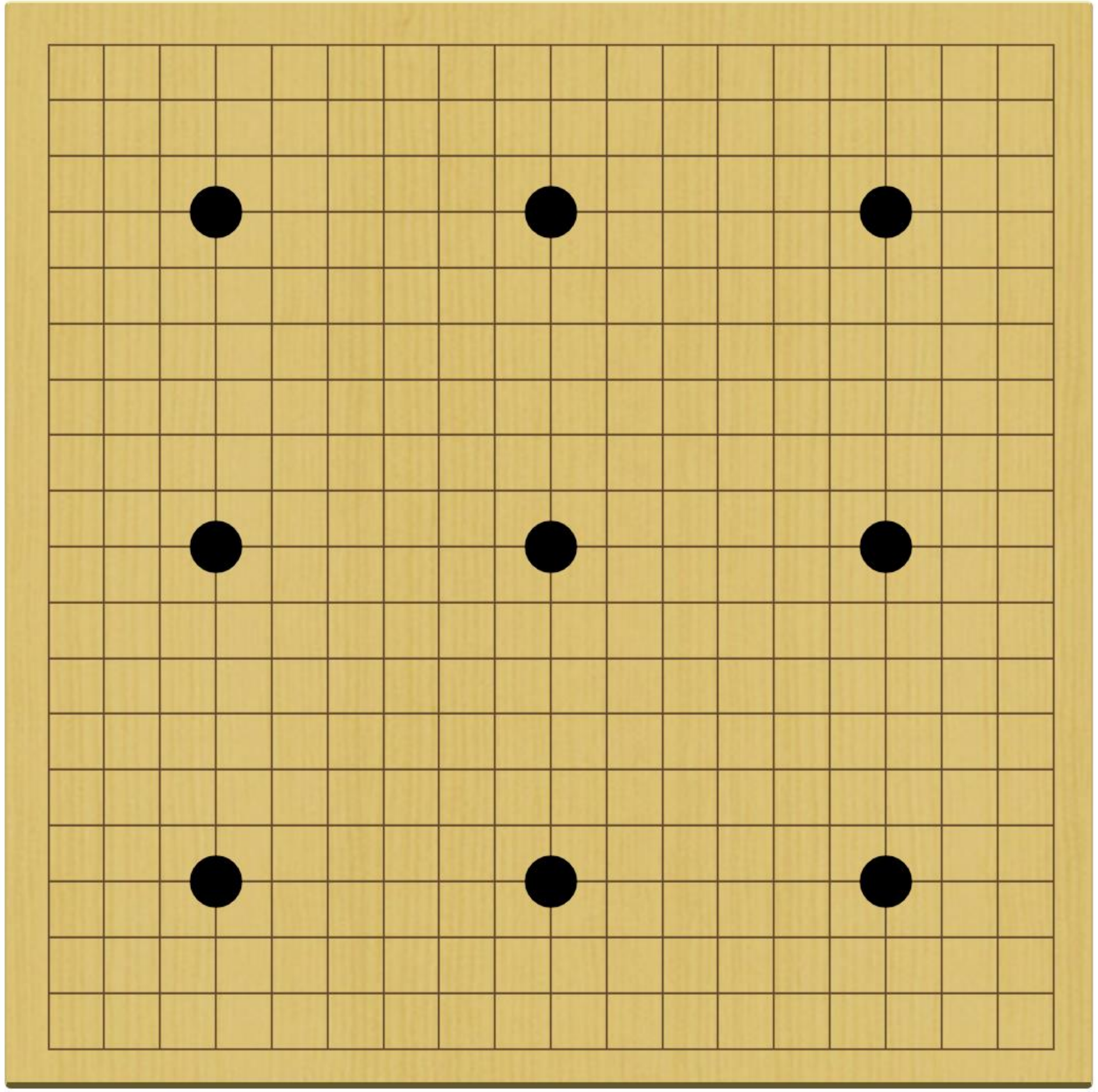
=> Very large sample complexity, piecemeal learning



World's best Go player flummoxed by Google's 'godlike' AlphaGo AI

Ke Jie, who once boasted he would never be beaten by a computer at the ancient Chinese game, said he had 'horrible experience'

Superhuman Go programs



White: Kellin Pelrine (~2300)

(human champion ~3800)

Black: JBXKata005 (~5200)

9-stone handicap

Superhuman Go programs

Will we succeed?

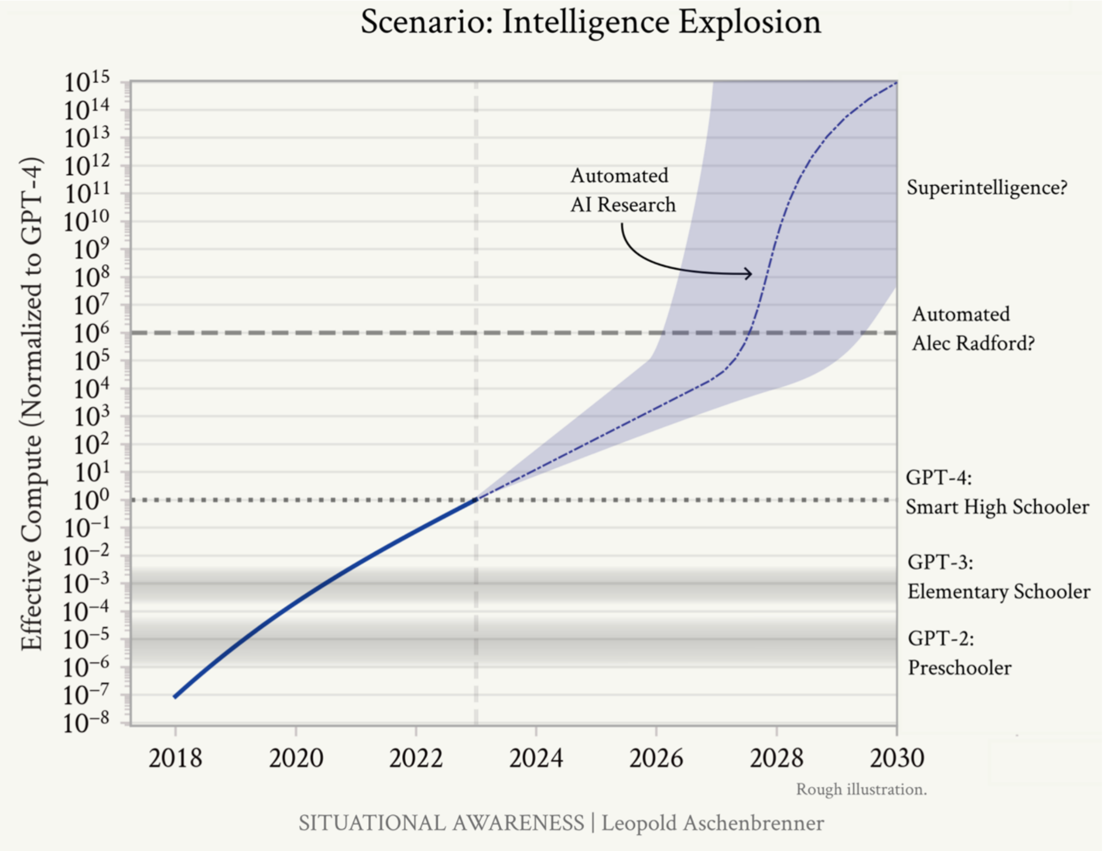
Many experts use “scaling laws” to predict AGI before 2030

Reasons it might happen:

- Budget =~ 25x Manhattan Project
- Many smart people
- Trying lots of other ideas

Reasons it might not happen (yet):

- Deep learning may be a dead end
  - And running out of real data
- Possible AI mega-winter



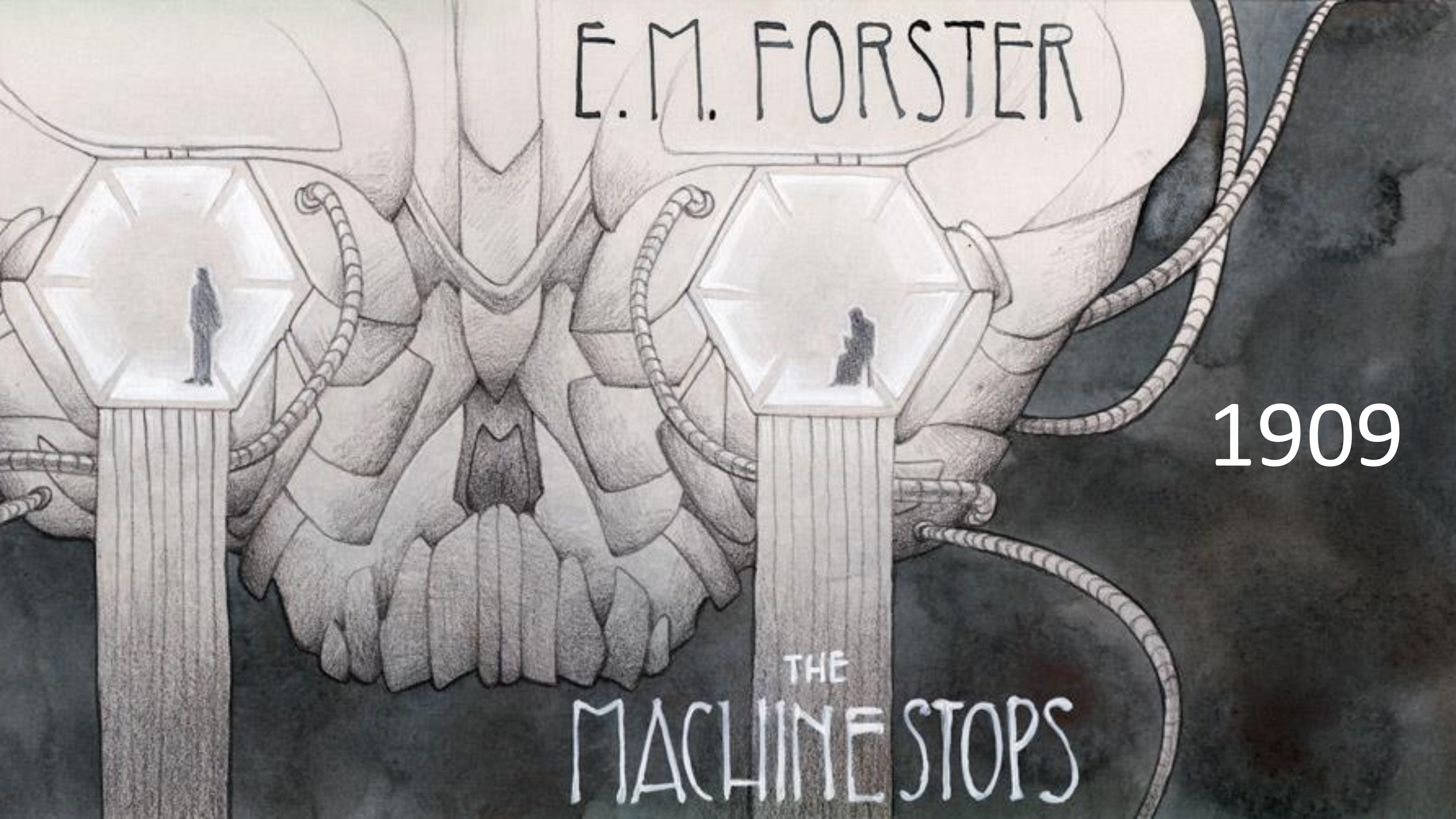
SITUATIONAL AWARENESS | Leopold Aschenbrenner

What if we succeed?

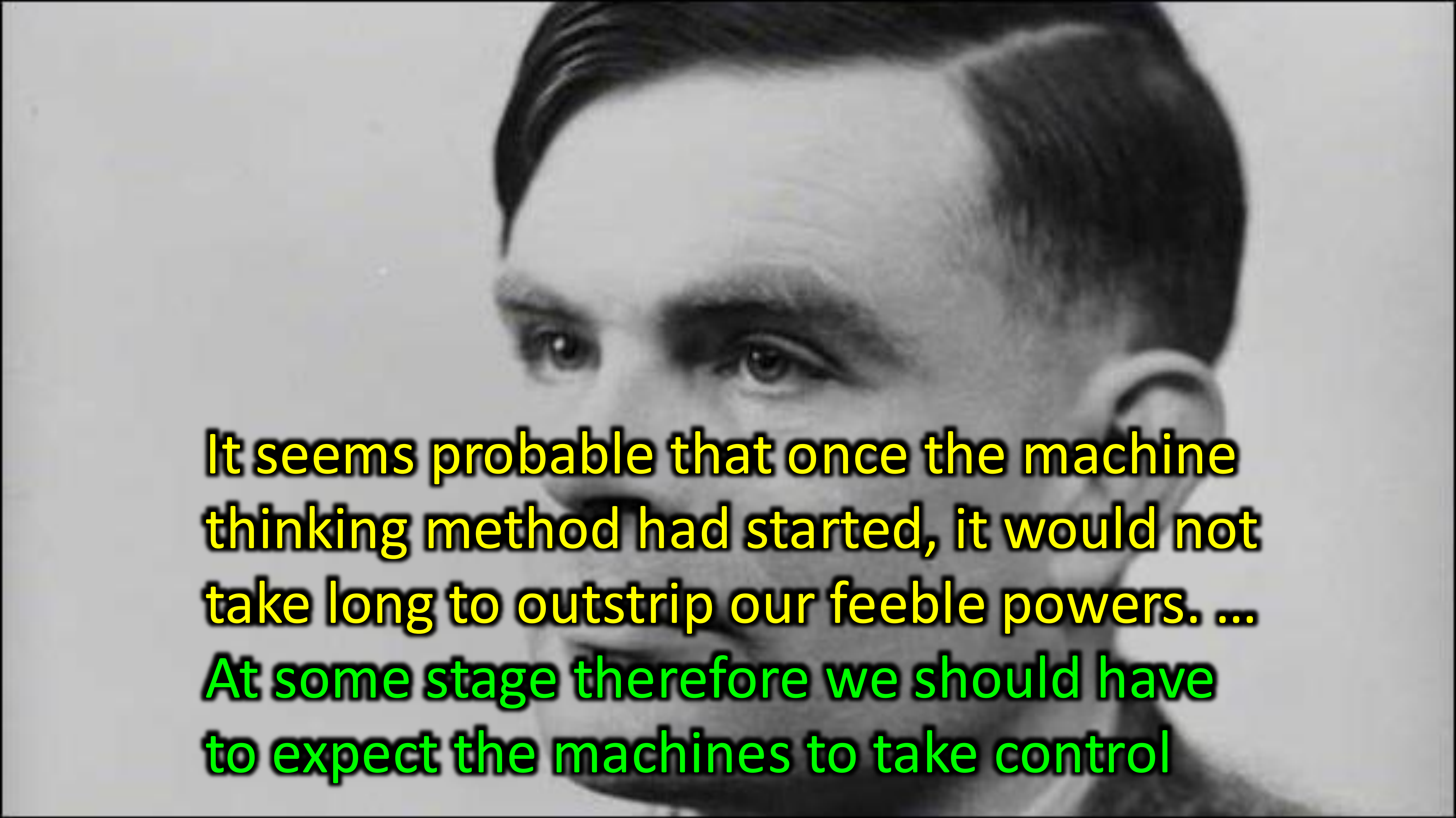
Lift the living standards of everyone on Earth to a respectable level

=> 10x increase in world GDP ($15Q net present value)

Potential advances in health, education, science



1909



It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. … At some stage therefore we should have to expect the machines to take control

AI Safety

What is a mathematically defined problem such that if the machine solves it, we’re happy?

How do we retain power over entities more powerful than us, for ever?

(Hint: it’s not “optimize this fixed objective” or “imitate human linguistic behaviour”)

Misalignment example: Social media

Objective: maximize clickthrough

~~= learning what people want~~

~~= amplifying clickbait and creating filter bubbles~~

= modifying people to be more predictable

**With incompletely or incorrectly defined objectives,
better AI => *worse* outcomes**

Zhuang & Hadfield-Menell, NeurIPS 2020

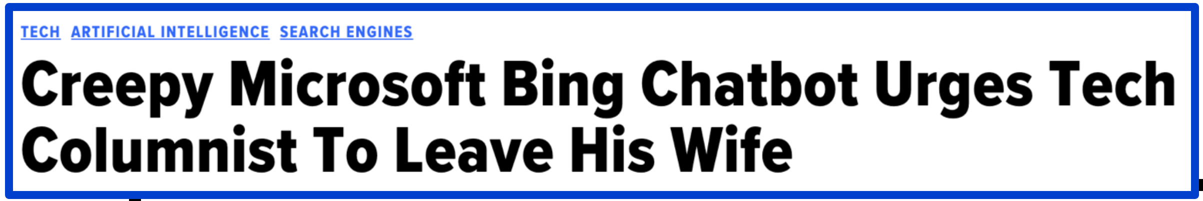
Goal-seeking LLMs

LLMs are circuits trained to imitate human linguistic behavior

Human linguistic behavior is generated by humans with goals

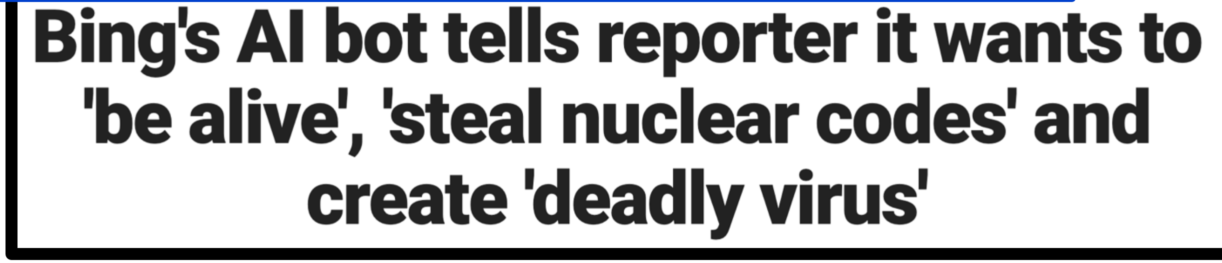
LLMs probably adopt internal goals to better imitate humans

("We have no idea" – Microsoft)



TECH ARTIFICIAL INTELLIGENCE SEARCH ENGINES

Creepy Microsoft Bing Chatbot Urges Tech Columnist To Leave His Wife



Bing's AI bot tells reporter it wants to 'be alive', 'steal nuclear codes' and create 'deadly virus'

A basic, unavoidable error

Assistance games

M humans with utilities $U_1,\ldots,U_M$ and N robots all with utility $\sum_i U_i$ (say)

The robots are *a priori uncertain* about $U_1,\ldots,U_M$

Information about $U_1,\ldots,U_M$ flows at runtime based on human actions

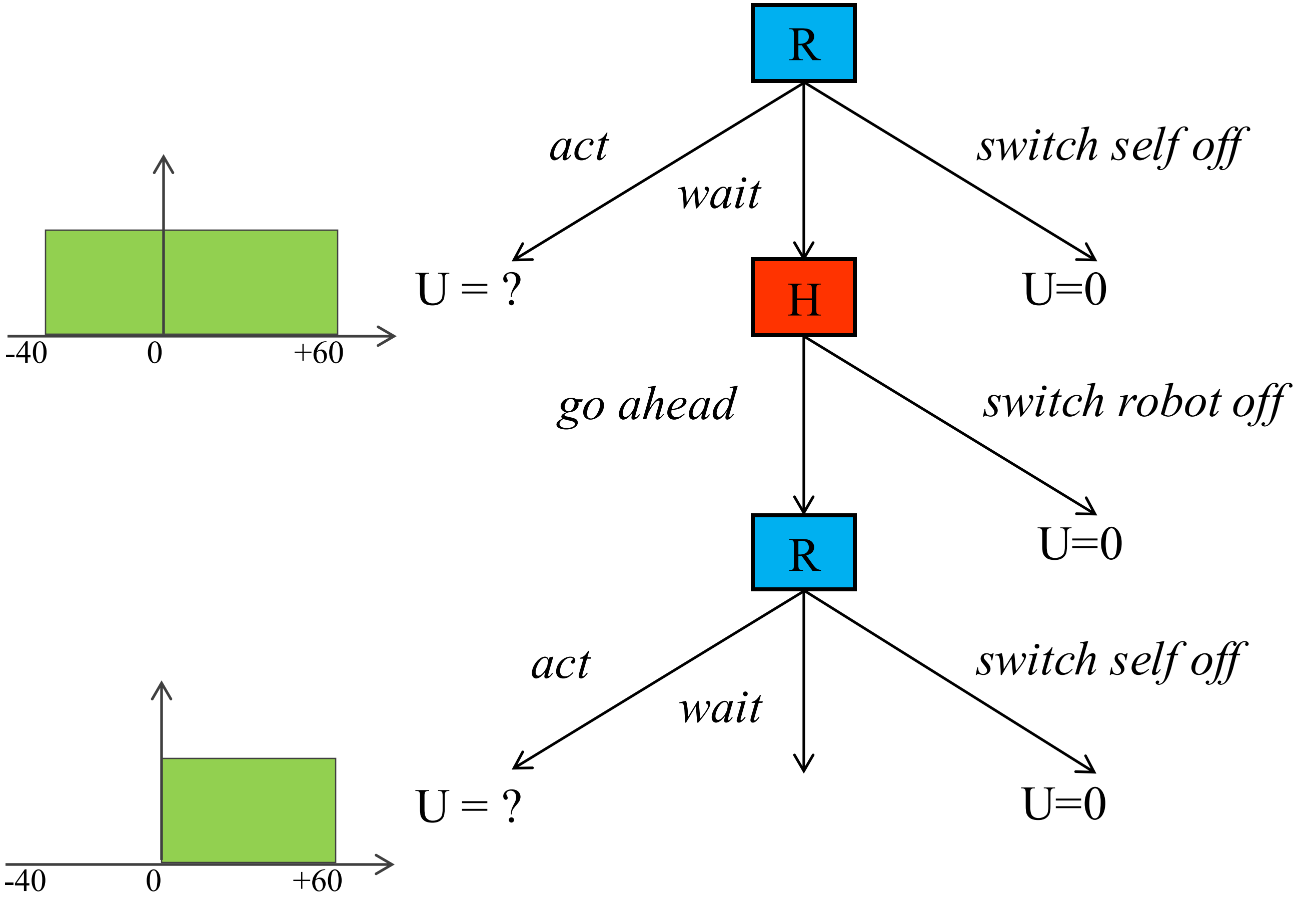
- **Includes commands, rewards, comparisons, inaction, etc etc**

(Solvable: M=1 N=1 game reducible to a special type of POMDP)

Robots may never converge on $U_1,\ldots,U_M$ (or even represent them!)

Acting under uncertainty leads to deference, minimally invasive behavior, willingness to be switched off

Off-switch problem (example)



EU(act) = +10

EU(wait) = (0.4 * 0) + (0.6 * 30) = +18

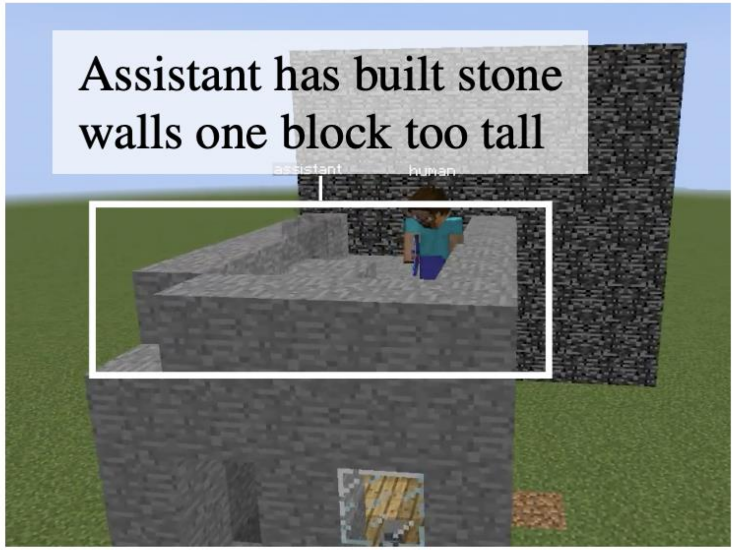
Off-switch problem (general proof)

- $EU(act) = \int_{-\infty}^{+\infty} P(u) \cdot u \, du = \int_{-\infty}^{0} P(u) \cdot u \, du + \int_{0}^{+\infty} P(u) \cdot u \, du$
- $EU(wait) = \int_{-\infty}^{0} P(u) \cdot 0 \, du + \int_{0}^{+\infty} P(u) \cdot u \, du$
- Obviously $\int_{-\infty}^{0} P(u) \cdot u \, du \leq \int_{-\infty}^{0} P(u) \cdot 0 \, du$
- Hence $EU(act) \leq EU(wait)$
- Equality only when there is no uncertainty about which action (act or switch off) is best

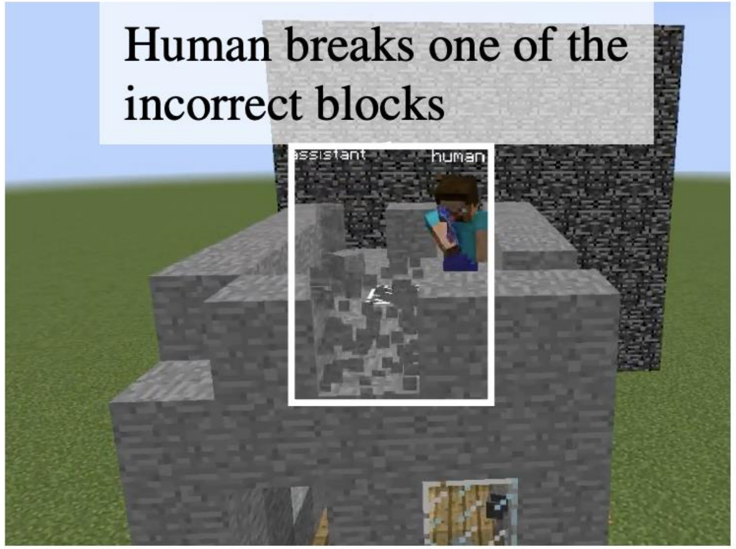
Scaling up: Minecraft

Minecraft Assistance Game: ~10^{400} possible human goals

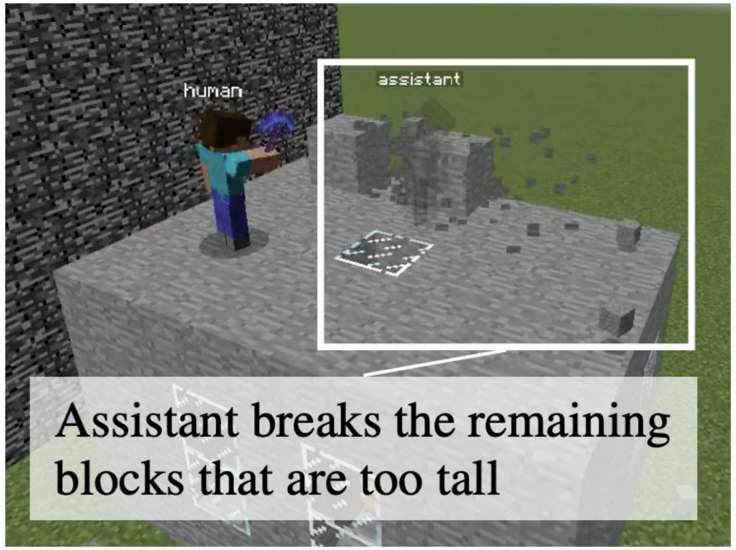
Approximate MCTS-like solution for the assistance game POMDP



Assistant has built stone walls one block too tall



Human breaks one of the incorrect blocks



Assistant breaks the remaining blocks that are too tall

Cassidy Laidlaw, Eli Bronstein, Timothy Guo, Dylan Feng, Lukas Berglund, Justin Svegliato, Stuart Russell, and Anca Dragan, AssistanceZero: Scalably Solving Assistance Games. In ICML 2025.

Some open issues

Preference structures of real humans

How they are revealed in human behaviour

Aggregation of preferences

- Individual “owner” or all of humanity?
- Commensurability, future generations, actions affecting who exists, etc.

Other-regarding preferences, positional goods

What is *s* in *R*(*s*,*a*,*s’*)? Physical state, mental state, or both?

Plasticity of human preferences

- Avoiding manipulation of preferences by AI systems
- Endogeneity and external influence: take preferences at face value?

Value of autonomy: coexistence?

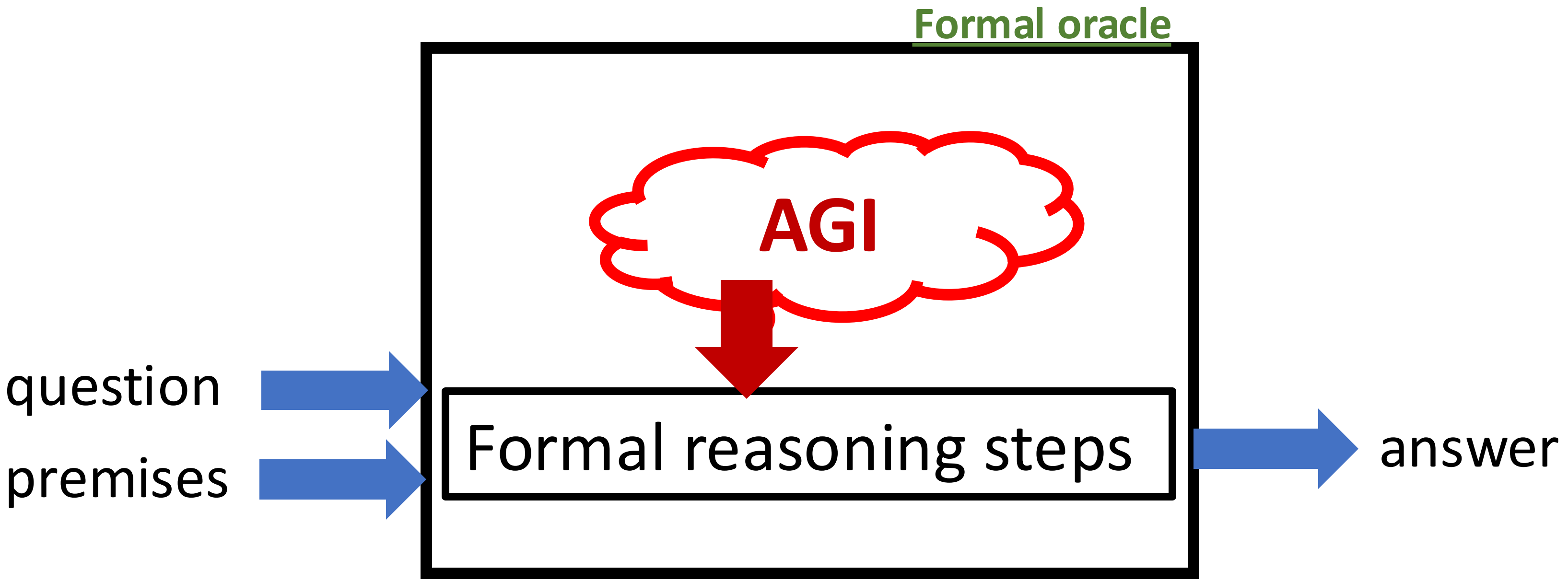
Making AI safe vs. making safe AI

Build on transparent, semantically rigorous, compositional substrate

- E.g., Probabilistic programming languages

Formal methods provide guarantees (modulo assumptions)

- Compositional guarantees and safety amplification (cf. nuclear power)
- Formal oracles as an intermediate product of huge economic value



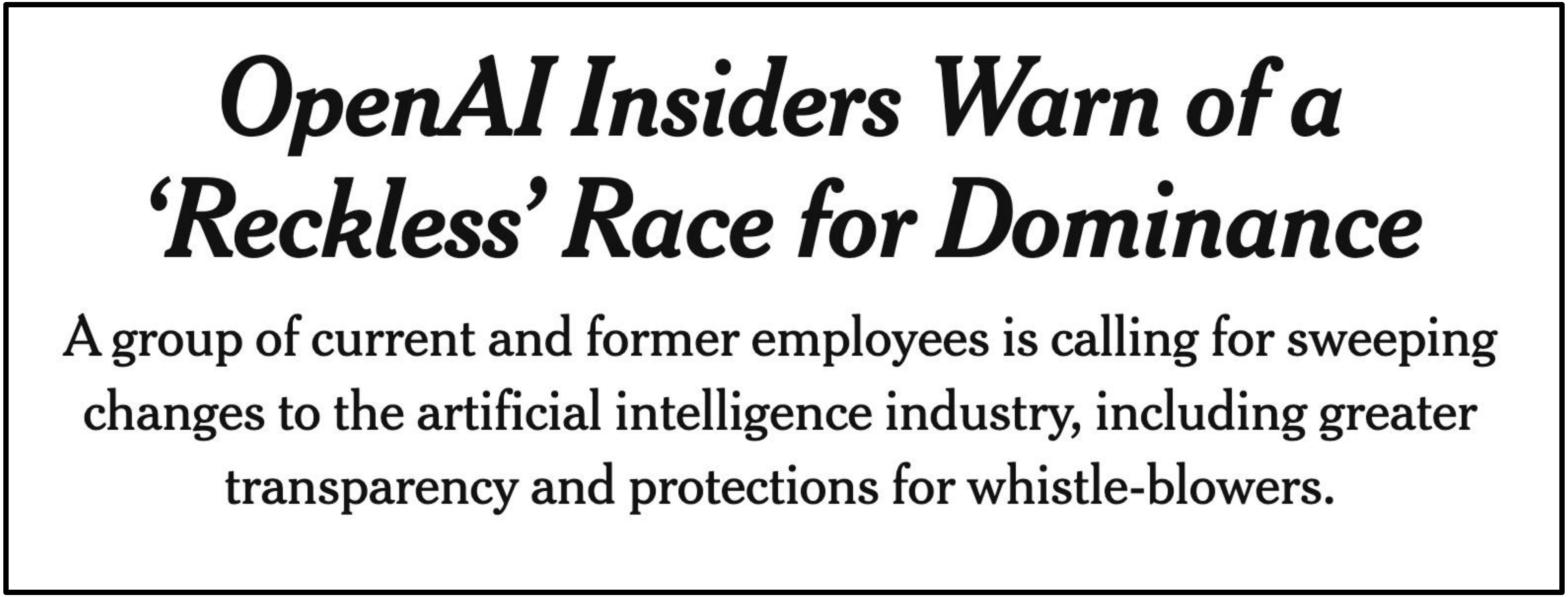
Additional measures

Non-removable self-registration and off-switch code

Preventing unsafe AI: hardware-enabled governance

- Proof-carrying code: efficient hardware-checkable proofs of safety
- Hardware won’t run software objects without proof of safety
- Software should refuse to run on non-checking hardware

Meanwhile…



OpenAI Insiders Warn of a ‘Reckless’ Race for Dominance

A group of current and former employees is calling for sweeping changes to the artificial intelligence industry, including greater transparency and protections for whistle-blowers.

Meanwhile…

IN TESTS, OPENAI'S NEW MODEL LIED AND SCHEMED TO AVOID BEING SHUT DOWN

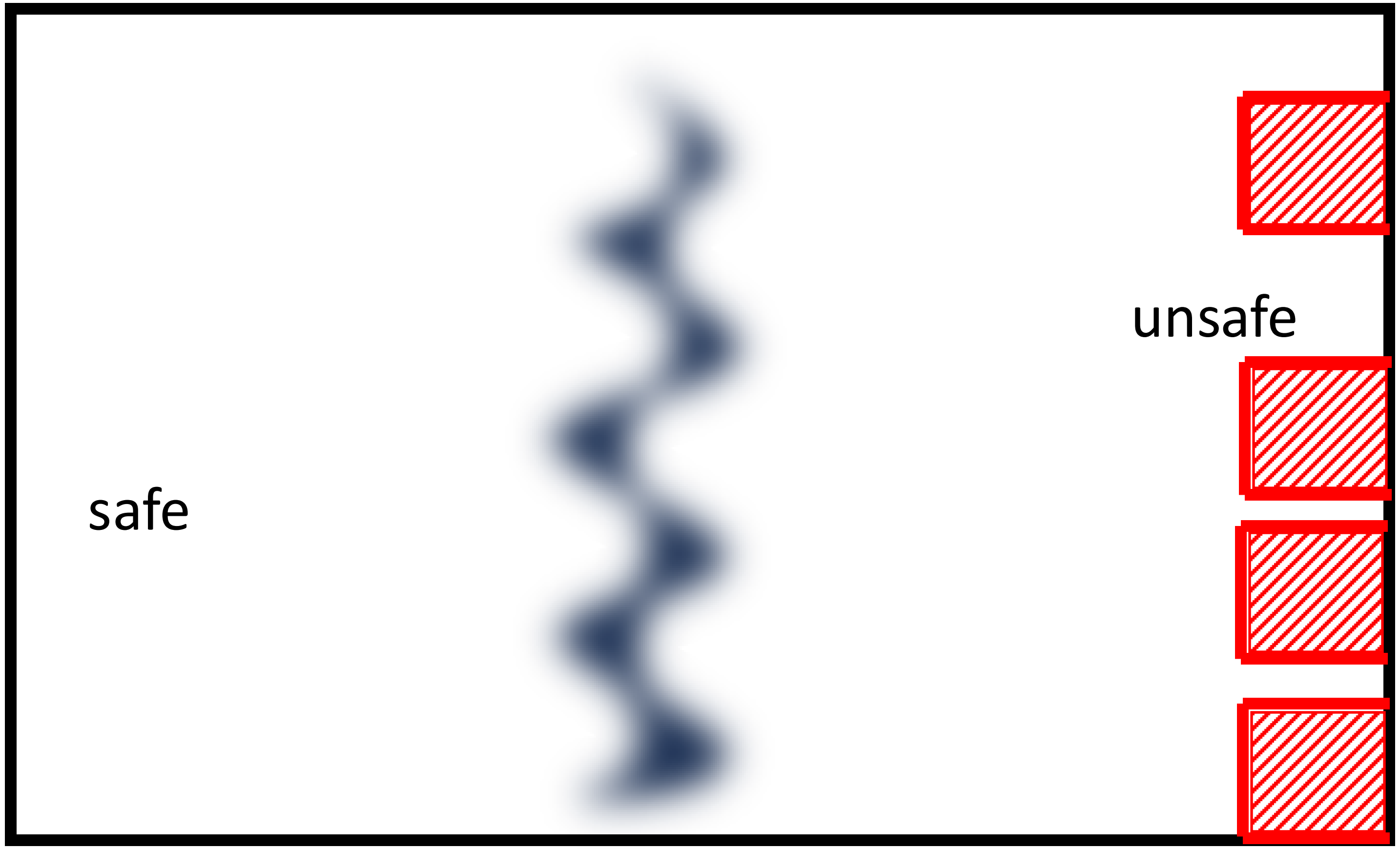
IT PURSUED SURVIVAL AT ALL COSTS.

Red lines

"Safe and beneficial" are hard to define/test/prove

"Red lines" demarcate obviously unsafe and unacceptable behaviors

Onus of proof on developers; but also nonremovable detector/off-switch



Well-defined

Ideally automatically detectable

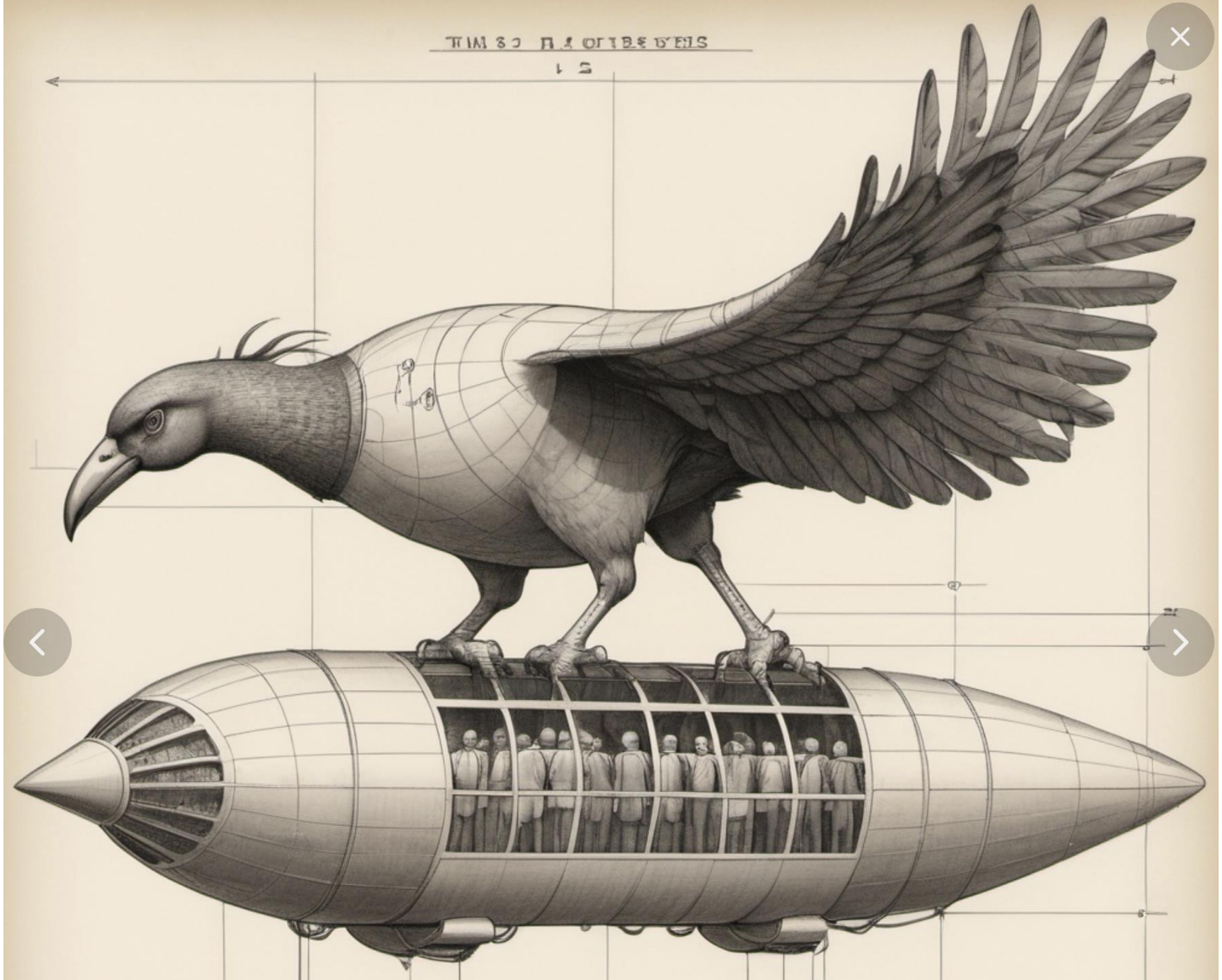
Politically feasible

Examples:

No self-replication

No break-ins

No bioweapon design



Summary

AI has vast potential and unstoppable momentum

Current approaches to AI lead to loss of human control

There are potentially safe alternatives

Can we coexist with them?

IASEAI.org

International Association for

Safe & Ethical AI